DOCUMENT RESUME

ED 167 605 TM 008 490

AUTHOR Lai, Morris K.
TITLE Consumer's Guide to Educational Evaluation.
PUB DATE 27 Mar 78
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Administrator Guides; *Educational Assessment; Educational Testing; Employment Qualifications; *Estimated Costs; Evaluation Methods; *Evaluation Needs; Evaluators; *Guidelines; Program Descriptions; *Program Evaluation; Testing Problems

ABSTRACT

Although much has been written about educational evaluation, few guidelines exist for consumers--project directors, school administrators, curriculum developers, legislators, teachers, parents, and boards of education. Several cautions surface from a review of the literature. First, tests that are based on program objectives are most useful to evaluations. Such objectives-based or criterion referenced tests often must be created by the evaluation staff, but test construction is difficult, lengthy, and costly. Second, goals or objectives themselves must be evaluated before program planning. Third, program design should be described in detail and fourth, consumers should know how to select qualified evaluators. Empirical cost data were collected on completed evaluations whose budgets ranged from $400 to $3 million. The data were studied to obtain estimates of costs expected in relation to sample size, number of schools involved, number of test items developed and used, report length, and project staff time. Although a more thorough investigation is needed, these preliminary results should furnish a basis for estimating the cost of a proposed evaluation or determining how much evaluation can be done for x amount of dollars. (CP)

CONSUMER'S GUIDE TO EDUCATIONAL EVALUATION

Morris K. Lai
Curriculum Research and Development Group
University of Hawai'i
27 March 1978

Paper presented at the annual meeting of the
American Educational Research Association
Toronto, Ontario, Canada

AERA Session # 4.18

CONSUMER'S GUIDE TO EDUCATIONAL EVALUATION

I. Introduction

Recently several books have been published on educational evaluation (e.g., Rippey, 1973; Worthen & Sanders, 1973; Anderson, Ball & Murphy, 1974; Borich, 1974; Popham, 1974; Walberg, 1974; Yelon, 1974; Cooley & Lohnes, 1976; Glass, 1976; Guttentag, 1977); however, very little has been written to assist the actual consumer of evaluation. This situation is understandable in light of the fact that contemporary educational evaluation is a relatively new discipline that many feel had its start with the publication of Scriven's 1967 paper entitled "The methodology of evaluation," or perhaps Cronbach's 1963 article, "Course improvement through evaluation." The evaluation writings that have followed, however, have been addressed mainly to practicing, aspiring, or forced-to-do-it evaluators, and little has been written to help the consumer of evaluation who does not actually do the evaluation. Included in this category of consumers are project directors, school administrators, curriculum developers, legislators, teachers, parents, and boards of education. Although the concerns of this paper are not directly addressed to evaluators, improved communication between consumer and evaluators requires that both groups be aware of the many problem areas.

In recognition of the fact that the typical consumer of evaluation has very little time to devote to evaluation matters, this paper will be kept as short as feasible. References are given with the realization that very few will be searched out and read; however, some consumers may want to go deeper into certain areas or be able to converse in a more informed manner with their evaluators. The overall goal of this paper is to assist the consumer in his/her understanding of what evaluators can or cannot do and how best to make use of evaluation.

II. State of the art of evaluation

A. Evaluation models and designs

As a relatively new field, evaluation has developed a jargon of its own. For the consumer who encounters technical evaluation terms and jargon, evaluation glossaries have been produced by Anderson, Ball, and Murphy (1973), the California Program Evaluation Improvement Project (1975), and Scriven and Roth (1977).

There is no standard approach to evaluation. Although several evaluation models or schemata have been proposed, there is virtually no evidence about the relative efficacy of the many models (Worthen, 1972; Smith & Murray, 1974). The consumer, however, need not blindly accept any model. Instead he/she can, as a start, look at Worthen and Sanders' (1973) summary of the strengths and weaknesses of eight of the major models* or ask the evaluator to explain why a given model was selected. Another fruitful endeavor would be for the consumer to look at the evaluation design or plan created by the evaluator. Criteria and guidelines to evaluating evaluations (and evaluation designs) have been presented by Stufflebeam et al. (1971), Scriven (1974), and Sanders and Nafziger (1975). The checklists by Scriven (1974) can be used to carry out evaluations as well as assist in meta-evaluations (the evaluation of evaluation plans). Ideally an evaluation plan would cover all of the 18 areas described by Scriven (needs, market, true field trials, true consumer, critical competitors, long term, side effects, process, causation, statistical significance, overall significance, costs, extended support). In actuality very few evaluations cover all of the areas adequately.

The evaluation plan or design is just the start of an arduous process that often brings into conflict the idealism of the plan and the realities of

* Those due to Alkin, Hammond, Personal Judgment, Provus, Scriven, Stake, Stufflebeam and Tyler.

operating in an educational environment. There are at least two major reasons why an evaluation design that is sound might nonetheless be followed by poor evaluation: 1) The measures used are invalid, 2) The educational environment interferes with the plans. In the first case, many consumers insist on the use of norm-referenced, standardized achievement tests. For most evaluations such tests are inappropriate because of invalidity in the form of overgenerality, nonoverlapping of objectives, bias, low reliability for individual test scores, questionable norms, confusing directions, and poor items. Organizations like the National Education Association (1973) and the National Advisory Committee on Mathematical Education (1975) have criticized the use of standardized testing.

B. Tests

A recent study (Lai, 1975) on the Stanford Achievement Test used in Hawai'i showed that the usual demographic (socioeconomic status) variables like family income or parents' education accounted for 86% of the variance in test scores, thus leaving very little that could be accounted for by, say, a program that was being evaluated. The consumer who, on the other hand, wants to know how a certain group of children is doing in relation to the rest of the United States is asking for a basically impossible comparison. If, for example, one is interested in comparing Hawai'i fourth graders with fourth graders in general, and he/she uses standardized achievement tests, then the comparison is invalid unless at least the following hold true: 1) Hawai'i's socioeconomic status is similar to that of the nation as a whole, or more correctly, that of the group used to establish the norms (there exists substantial evidence that this is not true), 2) the test does not measure objectives that are considered unworthy by a major group of consumers (this is important because for some subtests, only one or two

items can constitute a year's difference in grade equivalents[1]), 3) test administration was of the same caliber as that used when the norming was carried out (there exists evidence that this is generally untrue).

Tests that are based on program objectives are more useful to evaluations. Such objectives-based (similar terms are "criterion-referenced" or "domain-referenced") tests often have to be created by the evaluation staff. If the staff is inexperienced (e.g., has not previously developed good tests), then the consumer should not expect quality tests that are valid, reliable, easy to administer, clear in directions, etc. Test development is difficult, lengthy, costly (a discussion of cost guidelines is presented later in this paper), and not necessarily doable. If competence, time, or money is lacking, there is little hope. Preventive tactics on the part of the consumer could take the form of careful hiring of evaluators (see Section III) and/or ensuring that the program staff be prepared to state clearly what they would accept as evidence of success of a given program or product. Some idealists suggest that the test items be developed before any curriculum development takes place.

The generally low quality of tests available has been recognized in the literature (CSE, 1972, 1974), and some recommendations have been made. In areas like attitude toward math, for example, it has been recommended that reinventing the wheel be discouraged and use be made of what is already available (Aitken, 1976). Objectives banks (e.g., Instructional Objectives Exchange (IOX) or Westinghouse Learning Corporation) with associated items have been created in an attempt to help individuals select custom-made (based on objectives) measures. The consumer, however, is unlikely to be pleased with all of the items produced. Extensive evaluations of all pub-

[1] Problems associated with the use of grade equivalents have been written up elsewhere. In general consumers are advised to not use them.

lished tests have been made by the Center for the Study of Evaluation at UCLA (1972, 1974). A competent evaluator would be generally aware of tests available and be able to find appropriate ones if any. If test construction is required, the consumer must allow time for pilot tests of the measures. Because of the costs and time involved, it is usually best to keep the number of tests being developed to a minimum. It is better to get reasonably good data in a few areas than a lot of questionable data from many areas. Baker (1974) and Popham (1972) have advocated such an approach to evaluation which calls for lean (but important) data. The lean-data evaluation readily admits lack of comprehensiveness while at the same time defends the worthwhileness of collecting good data in a focused area. For some consumers and budgets, this approach may be the only feasible way to go.

Revisions of pilot versions of tests are carried out in the form of item analyses. It is rather easy for the consumer to be overwhelmed by the technical end of item analysis that uses terms like alpha coefficient, biserial correlations, factor analysis, or discrimination index. Oftentimes the program staff leaves too much in the hands of the external (outside of the program) test developer. In matters of test reliability or internal consistency, the evaluator/test developer is probably more qualified than the program staff; however, in terms of validity, a content analysis, preferably based on some theory, performed by the program staff is more likely to improve test validity than anything the evaluation staff does. Colwell (1970) and Guttman (1976) have emphasized this point in their writings.

Although Scriven (1972b) has advocated evaluation that is uncontaminated by program goals, his goal-free approach is infrequently used. The consumer should be sensitive to the possibility of side-effects (not directly related to program goals) which are often unanticipated. In the past, side-effects

have sometimes turned out to be major findings. (For example, the finding that tutors made substantial cognitive gains, whereas gains by tutees were minimal.)

C. Goals and objectives

A contemporary approach to evaluation requires that goals or objectives themselves be evaluated since it is of little value to know that goals were or were not attained if the goals themselves are not worthwhile ones. The evaluation literature has suggestions on ways to do this, and program staff are likely to be involved (e.g., see "Delphi technique" in the Scriven and Roth (1977) and Anderson et al. (1973) glossaries). This aspect is often intertwined with needs assessment in that goals should be related to needs. According to Stake (1970) objectives should be treated as fallible data. More attention should be given to statements of priority.

D. Program description

Good evaluation also requires that the product or program be described in some detail. It is not enough just to test participants before and after their exposure to a product or program. If a teaching process is involved, it should be observed and analyzed in terms of fidelity or quality of treatment. The modern approach to evaluation is strong on not taking for granted that program differences exist merely because there are two different labels (e.g., curriculum A and curriculum B). Charters and Jones (1973) have discussed the risks of appraising such non-events (no real differences in program presentation) in evaluation.

Evaluation must also take into account what characteristics the product or program users bring to the situation. Things like socioeconomic variables (e.g., family income, parents' education, ethnicity) and pretest cognitive

and affective (e.g., attitude) performance are often important to the interpretation of exit level performance. Where possible invasion of privacy is involved, the consumer and evaluator should be aware of the various Federal and local laws that apply (e.g., see Weinberger & Michael, 1977). Complex laws, of course, also apply to any contracts between consumer and evaluator.

III. Selecting the evaluator

All of the concerns discussed in Part II become painfully academic if an incompetent or lazy evaluator is hired. A few studies have been done on what skills evaluators need (e.g., Bunda (1973), Worthen (1975)), but the consumer who must hire an evaluator will not be in a good position to know if the desired skills are possessed by the evaluator. The best recourse is for the consumer to obtain, before hiring, examples of the evaluator's previous work. The consumer can then perform a quick metaevaluation using the criteria or approaches mentioned in Part II of this paper. It would also be useful to interview prospective evaluators to discuss their evaluation methods and philosophies. If the prospective evaluator is inexperienced in the on-the-job sense, then the academic training must be scrutinized. Evaluation specialization is offered at only a few universities or colleges. Phi Delta Kappa (1976) has published a comprehensive description of existing evaluation training programs. As a minimum an inexperienced evaluator should have had training in tests and measurements, statistics (preferably to the advanced levels of correlation, multivariate analysis, regression, etc.), research design, curriculum design, and evaluation. He/she should be a decent writer as well as a good communicator. Scriven (1972a) has gone so far as to suggest that a well-qualified evaluator is one who has all the skills known to humanity.

IV. Getting the evaluation done

The last obvious major pre-evaluation step is the contracting process. Again a checklist approach is available to help the consumer not only write up an evaluation contract but also decide whether or not the evaluation should even be done (Wright & Worthen, 1975).

In deciding how much money to set aside for evaluation, project directors and evaluators have had virtually no help from the literature. Rules of thumb like "5% and 10% of the program budget should be set aside for evaluation" have been given without much rationalization; furthermore, these rules of thumb have usually referred to desired amounts rather than minimal amounts needed to do a decent job.

A. Evaluation costs

In order to make more rational decisions about evaluation budgets, the decision maker needs to know what he/she can get for X amount of dollars. Here is where the professional evaluation literature, in general, is lacking. Empirical cost data collected on completed evaluations whose budgets ranged from $400 to $3 million were studied to obtain ball-park estimates of costs expected in relation to sample size, number of schools involved, number of instruments and items developed and used, report length, and project staff time. Because of the tremendous variation in the scope and costs of the sample of completed evaluations, it was not possible to come up with easily used categories that related evaluation parameters to costs.

Instead it was decided that representative examples might be of some use to evaluation consumers. A more in-depth study of evaluation costs is currently being undertaken and will be reported in a subsequent paper.

Some cautions are essential to the proper interpretation of the following figures. First of all, dollar amounts represent contracted costs, although

an attempt was also made to obtain information on "costs" in the form of labor or facilities provided by the funding agency for program being evaluated. Secondly, it was virtually impossible to compare the various evaluations in terms of amount of travel required or level of evaluation staff. Thirdly, no claims are made as to the representativeness of the sample. Although many types of evaluation organizations are represented, the attempt at sampling has been, at best, casual. Finally, where profit was involved (as opposed to evaluations done by non-profit organizations), such monetary results were not available. The one thing that all the included evaluations do have in common is the fact that they all represent completed evaluations for which contracted funds are known. Although a much more thorough investigation is needed in this area, the preliminary results presented here should be of some use as a guideline for evaluation consumers, many of whom have virtually no basis for estimating the cost of their proposed evaluation or determining how much evaluation they can get done for X amount of dollars. At the very least they will now have some ball-park figures based on actual contracted and completed evaluations. They will also see that evaluations are often rather expensive although Scriven has suggested that even very costly evaluations can, in essence, be cost-free if their benefits exceed their monetary costs.

CONTRACTED AMOUNT	TYPE OF CONTRACTOR (EVALUATOR)	PROGRAM BEING EVALUATED	DATE	DURATION (MONTHS)	SAMPLE SIZE						MEASURES (TESTS)	REPORT SIZE (PAGES)	COMMENTS
					STUDENTS	TEACHERS	PARENTS	CLASSES	SCHOOLS	DISTRICTS			
$400	Indiv.	Comp. Ed.	'77	1	200			7	5		2	15	Tests developed and administered by program staff
$900	Indiv.	Lang. Arts	'75	2									Modify and pilot test an attitude test
$4000	Indiv	Comp. Ed. ESAA	'77	6	2500				25		10	100	One measure per student. No test development. Tests administered by program staff
$6000	Two Indiv.	Medical Educ	'77	3	150						3	30	No test administration or development
$7500	Small Non Profit Org.	Bilingual Ed.	'77		407		100					200	Data collect by program staff
$20,000	Private Firm	Lang. Arts Math	'77	2				45	14		2	?	Observation/interview used. Instruments already developed

CONTRACTED AMOUNT	TYPE OF CONTRACTOR (EVALUATOR)	PROGRAM BEING EVALUATED	DATE	DURATION (MONTHS)	SAMPLE SIZE						MEASURES (TESTS)	REPORT SIZE (PAGES)	COMMENTS
					STUDENTS	TEACHERS	PARENTS	CLASSES	SCHOOLS	DISTRICTS			
$30,000	Corporation	Comp. Ed. Title I	'70 '71	12		62 studies						193	Analyzed evaluation studies
$37,000	Small Non-Profit Org	Lang. Arts	'77	12	900			32	31		12		Several measures revised or developed
$93,000	Regional Lab	Lang. Arts/ Math				1819	963	148					Tests administered by program staff
$99,000	Large Non-Profit Org.	Comp. Ed Title I	'71 '72	6		100 reports						275	Analysis of reports
$130,000	Large Non-Profit Org.	Comp. Ed. Title I	'74 '76	20		283 reports						110	Analysis of reports

CONTRACTED AMOUNT	TYPE OF CONTRACTOR (EVALUATOR)	PROGRAM BEING EVALUATED	DATE	DURATION (MONTHS)	SAMPLE SIZE						MEASURES (TESTS)	REPORT SIZE (PAGES)	COMMENTS
					STUDENTS	TEACHERS	PARENTS	CLASSES	SCHOOLS	DISTRICTS			
$232,000	Large Non-Profit Org.	Comp. Ed. Title I	'68-'71	18			42 sites				88 item questionnaire	1017	Nationwide
$1,680,000	Large Non-Profit Org.	Comp. Ed. Title I	'74-'77	29	3500		19 sites				Questionnaires	450	Nationwide
$3,[illegible]5,500	Large Non-Profit Org.	Comp. Ed. Title I Lang. Arts	'71-'76	60	50,000				222 [illegible] 29		Observation, MAT, STEP, Coop Reading	2100	Nationwide

B. Cost estimates

The original plan for collecting information on the costs of evaluations called for looking exclusively at completed contracts as opposed to proposal-type cost estimates. The rationale here was that consumers would benefit more from actual real-world figures than from proposed budgets which in actual bidding might turn out to be either at the low or high extremes. As it turns out, however, experience with completed evaluation contracts has enabled some institutions or persons to be skilled at estimating costs for certain types of evaluations. Because such estimates are also part of the real world in the sense that RFP's (Requests for proposals) often reflect those estimates, it was decided to include some discussion on the matter.

The Department of Education in the State of California has provided the following estimate for evaluation at the enumeration level in which data are collected on matters like funding history, program objectives, number of participants, budget, and other descriptive information. At this level no participant or other product data are collected. Evaluation activities include a) analysis of legislation, b) preparation, field testing, and distribution of forms, c) tabulation of responses, and d) report writeup. For a simple program with 1000 district responses, total cost for this type of in-house report would be around $15,000.

The Hawai'i State Department of Education has found that evaluations of programs used statewide (there are approximately 12,000 pupils per grade) have a minimal cost of about $100,000. Statewide Title I evaluations are expected to run in the $100,000 to $125,000 neighborhood.

V. Conclusion

Although this paper has not directly addressed the problem of reducing evaluation costs, it has laid the foundation for a subsequent effort to promote evaluation cost efficiency. What this paper hopes to reduce is the percent of evaluation consumers who are basically unknowledgeable about a) what evaluation can or cannot do and b) how to have evaluations done by another party. As consumers become more knowledgeable, they will be better able to make use of evaluation findings. In terms of Hutchinson's (1972) criteria for successfulness of evaluation, such an increase in evaluation use by the consumer would imply an increase in the quality of evaluation, which in turn can lead to an increase in the quality of the entities being evaluated.

REFERENCES

Aiken, L. R. Jr. Update on attitudes and other affective variables in learning mathematics. Review of Educational Research, 1976, 46 (2), 292-311.

Anderson, S. B., Ball, S., Murphy, R. T., & Associates. Enclyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1975.

Baker, E. L. The role of the evaluator in instructional development. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N. J.: Educational Technology Publications, 1974.

Borich, G. D. (Ed.). Evaluating educational programs and products. Englewood Cliffs, N. J.: Educational Technology Publications, 1974.

Bunda, M. A. A partial validation of the new conceptualization of evaluation competencies. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

CSE test evaluations. Los Angeles: Center for the Study of Evaluation, 1972, 1974.

California Program Evaluation Improvement Project. Glossary of terms related to educational evaluation. Sacramento: California State Department of Education, 1975.

Charters, W. W. Jr., & Jones, J. E. On the risk of appraising non-events in program evaluation. Educational Researcher, 1973, 2 (11), 5-7.

Colwell, R. The evaluation of music teaching and learning. Englewood Cliffs, N. J.: Prentice-Hall, 1970.

Conference Board of the Mathematical Sciences, National Advisory Committee on Mathematical Education. Overview and analysis of school mathematics, grades K-12. Washington, D. C.: Conference Board of the Mathematical Sciences, 1975.

Cooley, W. W., & Lohnes, P. R. Evaluation research in education. New York: Halsted, 1976.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Glass, G. V. (Ed.). Evaluation studies review annual, vol. 1, 1976. Beverly Hills: Sage, 1976.

Guttentag, M. (Ed.). Evaluation studies review annual, vol. 2, 1977. Beverly Hills: Sage, 1977.

Guttman, L. Workshop on non-metric analysis, Honolulu, December 1976.

Hutchinson, T. E. A critique of federal evaluative methodology: A review of what was done compared to what should be done. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Lai, M. K. Regression analyses of the relationship between socioeconomic variables and the Stanford Achievement Test and Otis-Lennon Test that were administered to Honolulu District fourth graders in Fall 1975. Honolulu: Hawai'i Curriculum Center, 1975.

National Education Association. Task force and other reports. Washington, D. C.: National Education Association, 1973.

Phi Delta Kappa. Evaluation training catalog. Bloomington, Indiana: Phi Delta Kappa, 1976.

Popham, W. J. Objectives-based management strategies for large educational systems. The Journal of Educational Research, 1972, 66 (1), 4-9.

Popham, W. J. (Ed.). Evaluation in education. Berkeley: McCutchan, 1974.

Rippey, R. M. (Ed.). Studies in transactional evaluation. Berkeley: McCutchan, 1973.

Sanders, J. R., & Nafziger, D. H. A basis for determining the adequacy of evaluation designs. Portland: Northwest Regional Educational Laboratory, 1975.

Scriven, M. The methodology of evaluation. In R. E. Stake (Ed.), Curriculum evaluation. AERA monograph series on curriculum evaluation, No. 1. Chicago: Rand-McNally, 1967.

Scriven, M. General strategies in evaluation. Curriculum Theory Network, 1972a, Monograph supplement, 182-192.

Scriven, M. Prose and cons about goal-free evaluation. Evaluation Comment, 1972b, 3 (4), 1-4.

Scriven, M. The checklist approach to formative evaluation and to product development. Mimeograph, 1974a.

Scriven, M. Evaluation perspectives and procedures. In W. J. Popham (Ed.), Evaluation in Education. Berkeley: McCutchan, 1974b.

Scriven, M., & Roth, J. Evaluation thesaurus. Berkeley: Scriven, 1977.

Smith, N. L., & Murray, S. L. The status of research on models of product development and evaluation. Northwest Regional Educational Laboratory (Research-Evaluation-Development Paper Series No. 1), 1974.

Stake, R. E. Objectives, priorities, and other judgment data. Review of Educational Research, 1970, 40 (2), 181-212.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. Educational evaluation and decision-making in education. Itasca, Ill.: Peacock, 1971.

Walberg, H. J. (Ed.). Evaluating educational performance. Berkeley: McCutchan, 1974.

Weinberger, J., & Michael, J. A. Federal restrictions on educational research: A status report on the privacy act. Educational Researcher, 1977, 6 (2), 5-8.

Worthen, B. R. Impediments to the practice of educational evaluation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Worthen, B. R. Competencies for educational research and evaluation. Educational Researcher, 1975, 4, 13-16.

Worthen, B. R. & Sanders, J. R. Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Wright, W. J., & Worthen, B. R. Standards and procedures for development and implementation of an evaluation contract. Northwest Regional Educational Laboratory, October, 1975.

Yelon, S. L. Constructive evaluation--Improving large scale instructional projects. Lansing: S. L. Yelon, 1974.